Title
Building fault detection and diagnostics: Achieved savings, and methods to evaluate algorithm performance

Authors
Lin, G
Kramer, H
Granderson, J

Publication Date
2020-01-15

DOI
10.1016/j.buildenv.2019.106505

Peer reviewed

DOI: 10.1016/j.buildenv.2019.106505

BERKELEY LAB

Lawrence Berkeley National Laboratory

Building fault detection and diagnostics: Achieved savings, and methods to evaluate algorithm performance

Guanjing Lin, Hannah Kramer, Jessica Granderson

Building Technology and Urban Systems Division
Lawrence Berkeley National Laboratory, Berkeley, CA

Energy Technologies Area
January 2020

Building Fault Detection and Diagnostics: Achieved Savings, and Methods to Evaluate Algorithm Performance

Guanjing Lin*, Hannah Kramer*, Jessica Granderson*

* Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 90R3111, Berkeley, CA, US 94720

Abstract

Fault detection and diagnosis (FDD) represents one of the most active areas of research and commercial product development in the buildings industry. This paper addresses two questions concerning FDD implementation and advancement 1) What are today's users of FDD saving and spending on the technology? 2) What methods and datasets can be used to evaluate and benchmark FDD algorithm performance? Relevant to the first question, 26 organizations that use FDD across a total 550 buildings and 97M sf achieved median savings of 8%. Twenty-seven FDD users reported that the median base cost for FDD software, annual recurring software cost, and annual labor cost were $8, $2.7 and $8 per monitoring point, with a median implementation size of approximately 1300 points. To address the second question, this paper describes a systematic methodology for evaluating the performance of FDD algorithms, curates an initial test dataset of air handling unit (AHU) system faults, and completes a trial to demonstrate the evaluation process on three sample FDD algorithms. The work provided a first step toward a standard evaluation of different FDD technologies. It showed the test methodology is indeed scalable and repeatable, provided an understanding of the types of insights that can be gained from algorithm performance testing, and highlighted the priorities for further expanding the test dataset.

Keywords: Fault detection and diagnostics; Energy efficiency; Savings and costs; Performance evaluation; Algorithm testing; Data

1. Introduction

Fault detection and diagnosis (FDD) is the process of identifying (detecting) deviations from normal or expected operation (faults) and resolving (diagnosing) the type of problem or its location. Automated FDD technologies can offer several interrelated benefits including energy savings and improved operational efficiency, utility cost savings, persistence in savings over time, streamlining operations and maintenance processes, and support for continuous energy management practices such as monitoring-based commissioning. The literature suggests that 5%-30% of commercial building energy is wasted due to problems associated with controls (Deshmukh 2018; Fernandez 2017; Granderson 2017a; Katipamula 2005; Roth 2005; Wall 2018).

While FDD has been in use in buildings for decades (Dexter and Pakanen 2001), its use is increasing, and today's market offers dozens of full-featured FDD software product offerings (Granderson 2017b, Smart Energy Analytics Campaign 2019a). These offerings integrate with building automation systems or can be implemented as retrofit add-ons to existing equipment, and continuously analyze operational data streams across many system types and configurations. This is in contrast to historically typical variants of FDD that are delivered as original equipment manufacturer-embedded equipment features, or those handheld FDD devices that rely upon temporary field measurements. With the upsurge in software, data availability, and data analytics across the buildings industry, new FDD algorithms are continuously being developed (Kim and Katipamula 2018, Lee 2019). A great diversity of techniques have been used for FDD, including physical models [Bonvini 2014, Muller 2013], black box [Zhao 2015, Zhao 2017], grey box [Sun 2014, Zogg 2016] , and rule-based approaches [Bruton 2014, House 2001]. Both the research and vendor communities are active in exploring new methods to improve the state of the art.

Although FDD is a powerful approach to ensuring efficient operations and the technology is maturing, it is still in the relatively early stage of adoption stock-wide. That is, in the language of technology adoption, today's users represent innovators and early adopters as opposed to early or late majority adopters. There is a wide range of questions that prospective users may confront as they consider whether or not to invest in implementing an FDD solution in their buildings. This paper addresses two specific questions from this broad spectrum of potential investigations:

1) What are today's users of FDD *saving and spending* on the technology?
2) What methods and datasets can be used to evaluate and benchmark FDD *technology performance?*

Relevant to the first question, we note that prospective FDD users must know the costs and savings of FDD technology to make a business case for technology investment and procurement decision-making. However, this information is only available in the literature in a limited number of case studies that document the savings and costs of commercially available FDD technology in real buildings. Summer (2012) reported annual energy cost savings of $18,400 and FDD installation costs of $94,500 at one building in the United States. Granderson (2017) showed 18.5% reduction in annual electricity consumption between 2009 and 2015 after Microsoft deployed the FDD-based Energy-Smart Buildings Program campus-wide. Wall (2018) indicated yearly savings of 8%-20% in electricity and 13%-28% in gas for FDD implementations at three buildings in the Austria.

To more comprehensively answer the question of FDD savings and costs, we draw from an ongoing public-private research partnership (Smart Energy Analytics Campaign 2019b) that engages analytics technology users to characterize and quantify the as-operated costs and benefits of technology use. We collected information from 36 organizations across the United States. These organizations use FDD in more than 200 million sq ft of commercial floor area and more than 2200 buildings. Each organization was asked to provide FDD technology costs, annual energy consumption before and after implementation of the FDD technology, and up to 10 of the most frequently implemented measures identified with support from their FDD from a list of 26 common operational improvements.

The second question is important given that prospective FDD users are notoriously challenged in distinguishing among the many FDD technology offerings on the market – particularly when it comes to knowing whether a given tool's underlying algorithm is sound, or any better performing than another's. In the field of building analytics software, prior work has established methods for testing the accuracy of building simulation software (ASHRAE 2014) and the energy information software that uses models for automated savings estimation (Granderson and Price, 2014, Granderson et al. 2015). Specific to fault diagnostics, while numerous research papers evaluate the performance of individual algorithms (Rossi and Braun 1997, Katipamula et al. 1999, Ferrettu et al. 2015) it is difficult to draw comparisons or understand the overall state of technology, as each study uses different datasets, test conditions, and metrics. A body of work by Yuill and Braun has explored these concerns, largely with a focus on handheld FDD devices for use with unitary systems (Yuill and Braun 2013, Yuill and Braun 2016, Yuill and Braun 2017). There is a lack of standard methodology and datasets for evaluating the accuracy of FDD technologies that continuously analyze operational data streams from building automation systems and built-up as well as unitary HVAC systems. In response, we describe a previously developed methodology for evaluating the performance of FDD algorithms (Frank et al. 2018, 2019a; Yuill and Braun 2013), and a newly curated initial test dataset of AHU system faults, with known ground-truth conditions. We've applied the evaluation methodology on three sample FDD algorithms, including two commercial tools, and an instantiation of National Institute of Standards and Technology's (NIST's) air-handling unit performance assessment (APAR) rules (House et al. 2001) against the dataset to understand the types of performance insights that can be gained, priorities for further expanding the test dataset for maximum utility in evaluating FDD algorithm performance, and whether the test methodology is scalable and repeatable.

In summary, this paper provides two primary contributions to existing work and help organizations adopt the FDD technology. 1) It quantifies the achieved savings and costs of FDD use over a large cohort of users, using a consistent study design. This moves beyond one-off case studies to provide a more complete picture of the FDD value proposition, based on data from field implementations. While this type of analysis has been conducted for meter-analytics and visualization technologies (Granderson 2016), it has not been done for fault detection and diagnostics technologies. 2) It provides a first step toward a standard evaluation of different FDD technology and algorithms, including a general methodology, a newly curated initial test dataset of AHU system faults, and a trial to demonstrate the process on two commercial FDD tools and a research-grade FDD algorithm. In the long term, this public dataset will be expanded. Once a comprehensive dataset is curated across a wide diversity of systems and equipment, it will be possible to benchmark the state of the art, supporting FDD researchers and developers improve algorithms, and potentially to enable standard certification processes.
In the remainder of this paper we describe the research methodology, followed by the results and a discussion of the findings for the two research questions respectively. The final section presents conclusions and future work.

2. Methodology

2.1 Assessment of FDD benefits and costs

To address the first research question concerning FDD costs and savings, we collected data from 36 geographically diverse US-located organizations using FDD technology. The data included basic building and technology information, year-over-year energy use trends, most frequently implemented energy efficiency measures, and technology costs. As opposed to equipment-embedded 'on-board' diagnostics, or other flavors of FDD, these users have implemented FDD solutions that comprise software systems that continuously integrate operational data from the building automation system (BAS) as well as stand-alone meters and sensors meters. These full-featured software solutions commonly contain large libraries of automated fault detection logic that span multiple systems, subsystems, and components (Granderson et al. 2017b). The study cohort comprised offices and buildings from the higher education market sectors, representing 2200 buildings and over 200 million square feet of floor area. The results of the assessment are presented in section 3.1.

Organizations were asked to provide annual energy consumption before and after FDD implementation. Energy savings since installation of the FDD were determined in two ways, one with interval data (hourly to 15-minute), and the other with monthly bill data, reflecting the diversity of savings analysis approaches in this study.

> Method 1. Interval data analysis: Pre-FDD (baseline year) interval data was used to develop a model of building energy use. Energy use was projected using the baseline model and compared actual energy use during the period after installing the FDD. This method utilizes the IPMVP Option C methodology (EVO 2016).
>
> Method 2. Monthly bill analysis: Pre-FDD (baseline year) energy use based on annualized monthly bills was compared to the most recent full year of energy use. Where possible, the data was normalized for weather using ENERGY STAR Portfolio manager (ENERGY STAR 2019).

Although occupancy rates can be a key driver of building energy consumption, the participating organizations were not able to provide occupancy data across the thousands of buildings that were included in the study. The effect of fluctuations in occupancy is not able to be controlled for in the study, however, this effect is mitigated by analyzing savings at the portfolio level for each participating organization. For all but two cases, respondents provided the researchers annual energy consumption, and the researchers calculated savings by comparing the annual energy use before and after FDD implementation (method 2 above). In the remaining two cases, the respondents provided the researchers a calculation of savings based on method 1 above. Energy cost savings are calculated using national average energy prices.

To further understand how FDD technology enables the savings achieved, organizations were asked to indicate up to ten of the most frequently implemented measures that they identified using their FDD from a list of 26 common operational improvement opportunities as shown in Table 1.

Costs to implement and use the FDD technology were gathered from study participants in the following categories:

- Base cost: Cost for FDD software installation and configuration, including FDD vendor and service provider costs. It does not include additional costs such as the cost of energy metering hardware and communications, adding points to the BAS, or retrofits.
- Recurring cost: Recurring annual cost for software license or software-as-a-service fees.
- In-house labor cost: Cost was determined using estimated hours for the team and $125/hour as average labor rate. The estimated hours are the approximate time spent by in-house staff reviewing FDD outputs, identifying opportunities for improvement, and implementing measures.

Table 1 Twenty-six common operational improvement measures.

Category	Specific Measure
Scheduling Equipment Loads	**Improve scheduling for HVAC & Refrigeration**: shorten operating hours of HVAC & refrigeration systems to better reflect actual building occupancy schedule and service needs.
	Improve scheduling for lighting: minimize the lighting runtimes.
	Improve scheduling for plug loads: minimize office equipment runtimes, e.g. installing advanced power strips which automatically cut power according to an occupant-defined schedule.
Economizer/Outside Air Loads	**Improve economizer operation/use**: repair/optimize the mixed air economizer control in an AHU (e.g., fix dampers, replace damper actuators, modify economizer control sequence, etc.).
	Reduce over-ventilation: adjust the minimum outdoor air ventilation setpoint to reduce heating and cooling loads.
Control Problems	**Reduce simultaneous heating and cooling**: eliminate unintended simultaneous heating and cooling by repairing the stuck/leaking coil valve, sensor errors, etc.
	Tune control loops to avoid hunting: adjust equipment/actuator controls to reduce cycling (turning on and off).
	Optimize equipment staging: add or optimize the equipment staging control (i.e., turning the equipment on to meet the load while maintaining optimum part-load performance)
	Zone rebalancing: ensure proper airflow to be delivered to each zone.
Controls: Setpoint Changes	**Adjustment of heating/cooling and occupied/unoccupied space temperature setpoints:** add or optimize controls of the zone terminal units to allow spaces temperatures to drift more during occupied/unoccupied hours.
	Reduction of VAV box minimum setpoint: reduce the VAV box minimum setpoint to reduce the heating and cooling load.
	Duct static pressure setpoint change: reduce the duct static pressure setpoint to reduce fan energy consumption.
	Hydronic differential pressure setpoint change: reduce the hydronic differential pressure setpoint to reduce pump energy consumption.
	Preheat temperature setpoint change: reduce AHU preheating settings.
Controls: Reset Schedule Addition or Modification	**Supply air temperature reset**: add or optimize control of the supply air temperature based on either outside air temperature or space loads.
	Duct static pressure reset: add or optimize control of the duct static pressure based on either outside air temperature or space loads.
	Chilled water supply temperature reset: add or optimize control of the chilled water supply temperature based on either outside air temperature or cooling load.
	Hot water supply temperature reset or hot water plant lockout: add or optimize control of the hot water supply temperature based on either outside air temperature or heating load.
	Condenser water supply temperature reset: add or optimize control of the condenser water supply temperature based on either outside air wet-bulb temperature or chiller load.
Equipment Efficiency Improvements	**Add or optimize variable frequency drives (VFDs)**: add a VFD to the fan or pump.
	Pump discharge throttled or over-pumping and low delta T: fix pump issues to allow it provide the proper water flow.
Occupant Behavior Modification	**Routinely share energy information or guidance on proper use of equipment with occupants through FDD technology**
	Hold an energy savings challenge using FDD data
Retrofits	**Lighting upgrade or improve lighting controls**: replace lighting fixtures with more efficient fixtures, add lighting control system.
	High efficiency HVAC equipment: airside: replace airside HVAC equipment with more efficient equipment.
	High efficiency HVAC equipment: waterside: replace waterside HVAC equipment with more efficient equipment.

2.2 FDD algorithm evaluation methodology and dataset

To address the second research question concerning methods and datasets to evaluate and benchmark FDD, we developed a general FDD algorithm performance evaluation methodology, curated an initial test dataset of AHU system faults, and completed a trial to demonstrate the evaluation process on two commercial FDD tools and a research-grade FDD algorithm.

The general FDD algorithm performance evaluation methodology is illustrated in six steps in Figure 1, with the procedure documented in Yuill and Braun (2013) as a starting point. Components 1, 2, 4, and 5 are original to the evaluation procedure presented by Yuill and Braun (2013), while components 3 and 6 have been added for clarity in implementation and execution.

1. Determine a set of input scenarios, which define the driving conditions, fault types, and fault intensities (fault severity with respect to measurable quantities).
2. Create a set of input samples drawn from the input scenarios, each of which is a test data set for which the performance evaluation will produce a single outcome.
3. Assign ground truth information to each input sample, e.g. faulted or unfaulted, and if faulted, which fault cause is present.
4. Execute the FDD algorithm that is being evaluated for each input sample. The FDD algorithm receives input samples and produces fault detection and fault diagnosis outputs
5. Retrieve FDD algorithm outputs (fault detection and diagnosis results).
6. Evaluate FDD performance metrics. First, raw outcomes are generated by comparing the FDD algorithm output and the ground truth information for each sample. Then, the raw outcomes are aggregated to produce performance metrics.

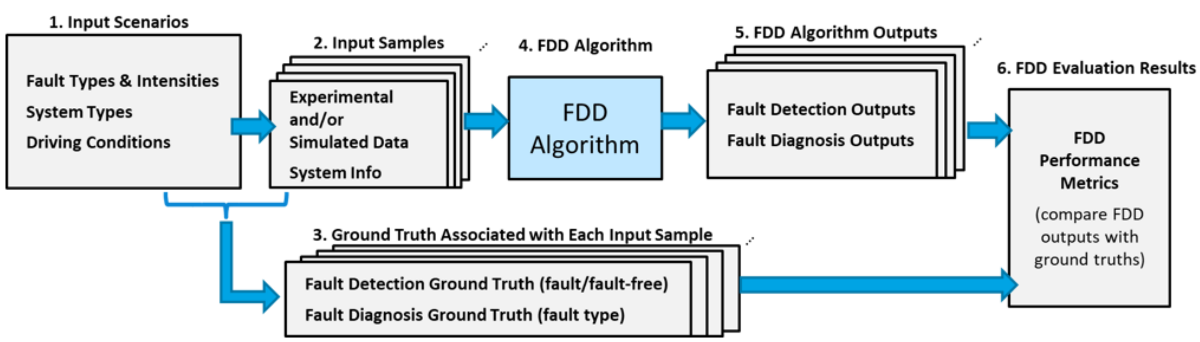

Figure 1. Automated FDD performance evaluation procedure, generalized and adapted from (Yuill and Braun 2013)

Frank et al. 2018 further documents options to define the input samples, ground truth fault conventions, and performance metrics. In the work presented in this paper, the evaluation methodology was applied with a preliminary dataset using a condition-based ground truth convention, and daily input samples. Metrics used for detection performance included false positive and false negative rate, true positive and true negative rate, and no detection rate. For diagnostic performance we used the correct, misdiagnosis, and no diagnosis rate. These terms and metrics are defined in Equations 1-7.

True positive refers to the case in which the ground truth indicates a fault exists and the algorithm correctly reports the presence of the fault.

The true positive rate, $TPR = \frac{\#\ of\ input\ samples\ with\ true\ positive}{\#\ of\ faulted\ input\ samples\ provided\ to\ FDD\ algorithm}$ (1)

True negative refers to the case in which the ground truth indicates a unfaulted state and the algorithm correctly reports a unfaulted state.

The true negative rate, $TNR = \frac{\#\ of\ input\ samples\ with\ true\ negative}{\#\ of\ unfaulted\ input\ samples\ provided\ to\ FDD\ algorithm}$ (2)

False positive refers to the case in which the ground truth indicates a unfaulted state but the algorithm reports the presence of a fault. It is also known as a false alarm or Type I error.

The false positive rate, $FPR = \frac{\#\ of\ input\ samples\ with\ false\ positive}{\#\ of\ unfaulted\ input\ samples\ provided\ to\ FDD\ algorithm}$ (3)

False negative refers to the case in which the ground truth indicates a fault exists but the algorithm reports an unfaulted state. It is also known as missed detection or Type II error.

The false negative rate, $FNR = \frac{\#\ of\ input\ samples\ with\ false\ negative}{\#\ of\ faulted\ input\ samples\ provided\ to\ FDD\ algorithm}$ (4)

No detection refers to the case in which the algorithm cannot be applied (for example, due to insufficient data) or the algorithm gives no response because of excessive uncertainty.

The no detection rate, $NDR = \frac{\#\ of\ input\ samples\ with\ no\ detection}{\#\ of\ input\ samples\ provided\ to\ FDD\ algorithm}$ (5)

Correct diagnosis refers to the case in which the predicted fault type (cause) reported by the algorithm matches the true fault type defined in the ground truth.

The correct diagnosis rate, $CDR = \frac{\#\ of\ input\ samples\ with\ correct\ diagnosis}{\#\ of\ faulted\ input\ samples\ provided\ to\ FDD\ algorithm}$ (6)

Misdiagnosis refers to the case in which the predicted fault type does not match the true fault type defined in the ground truth.

The misdiagnosis rate, $MDR = \frac{\#\ of\ input\ samples\ with\ misiagnosis}{\#\ of\ faulted\ input\ samples\ provided\ to\ FDD\ algorithm}$

No diagnosis refers to a case in which the algorithm does not or cannot provide a predicted fault type, for example, because of excessive uncertainty.

The no diagnosis rate, $NDgR = \frac{\#\ of\ input\ samples\ with\ no\ diagnosis}{\#\ of\ faulted\ input\ samples\ provided\ to\ FDD\ algorithm}$ (7)

The newly curated initial test dataset of AHU system faults consists of five groups of dataset (Table 2). The ground truth dataset for AHU faults was created using experimental test facilities as well as simulation models. The test facilities included Lawrence Berkeley National Laboratory's FLEXLAB™ (ETA 2019) and the Energy Resource Station at the Iowa Energy Center (Wen and Li 2011). The simulation models comprised a Modelica (LBNL 2018) representation of a multi-zone AHU-VAV system and HVACSim+ (Wen and Li 2011) representations of a multi-zone AHU-VAV system. Operational data for 75 24-hr periods of fault-free (28 days) and fault-present (47 days) conditions were collected (US DOE OpenEI), as summarized in Table 2.

Faults were imposed one at a time (that is, no test case comprised multiple faults), for a minimum of one day at each fault-intensity. The measurement points included in the dataset are representative of points commonly monitored in building control systems. Measured at a 1-minute frequency these points included:

AHU: Supply Air Temp. (°F) AHU: Supply Air Temp. Setpoint (°F)

AHU: Outdoor Air Temp. (°F) AHU: Mixed Air Temp. (°F)

AHU: Return Air Temp. (°F)	Occupancy mode (1-occupied, 0-unoccupied)
AHU: Supply Air Fan Status (1-on, 0-off)	AHU: Return Air Fan Status (1-on, 0-off)
AHU: Supply Air Fan Speed Control Signal (0-1)	AHU: Return Air Fan Speed Control Signal (0-1)
AHU: Outdoor Air Damper Control Signal (0-1)	AHU: Return Air Damper Control Signal (0-1)
AHU: Cooling Coil Valve Control Signal (0-1)	AHU: Heating Coil Valve Control Signal (0-1)
AHU: Supply Air Duct Static Pressure Set Point (psi)	AHU: Supply Air Duct Static Pressure (psi)

Table 2. Summary of the initial test dataset for AHU faults. The number in each cell indicates the number of 24-hour periods for which data was obtained for each fault scenario.

Input Scenarios			MZVAV AHU-1 (Sim.)	MZVAV AHU-2 (Exp.)		MZVAV AHU-2 (Sim.)			SZCAV AHU (Exp.)	SZVAV AHU (Exp.)
Fault type		**Fault**	**Spring**	**Spring**	**Summer**	**Spring**	**Summer**	**Winter**	**Winter**	**Summer**
OA damper	Stuck	Min. position				1		1		1
		Fully open							1	1
		40% open				1				
		45% open					1			
		50% open					1		1	
Valve of Heating Coil	Stuck	Fully closed							1	
		Fully open							1	1
		50% open							1	1
	Leakin g	Low			1		1		1	
		Medium			1		1			
		High			1		1		1	1
Valve of Cooling Coil	Stuck	Fully closed				1			1	
		Fully open				1	1		1	1
		15% open					1			
		50% open							1	
		65% open					1			
	Leakin g	Low							1	
		High							1	1
Outdoor air temp. sensor	Bias	+4F	6							
		-4F	6							
Unfaulted			6	4	3	3	9	1	1	1

To execute a trial of the evaluation methodology FDD algorithm developers were provided a description of the HVAC system including its type, a schematic diagram and the associated control sequences, and a list of the measurement points included in the dataset. They were not provided the ground truth information specifying which faults were present on which days in the dataset.

In the trial to demonstrate the evaluation process, two commercial FDD tools and a research-grade FDD algorithm were selected. The research-grade FDD algorithm is an instantiation of National Institute of Standards and Technology's (NIST's) air-handling unit performance assessment (APAR) rules (House et al. 2001). The two commercial FDD tools are two common tools that are used by several of the organizations in the study cohort for FDD savings and costs. Following the steps in the evaluation methodology (see Figure 1), the FDD algorithms were executed against the input samples, and the research team directly compared the algorithm output to the ground truth information. The algorithm outputs for each input sample were collated to calculate the performance metrics. The results of trial are presented in the section 3.2.

3. Results

Results for the benefits of FDD use are presented, followed by results from trialing the FDD algorithm evaluation test dataset and methodology.

3.1 As-operated FDD benefits and costs

To understand the benefits and costs of FDD to users of the technology, three primary indicators were considered. These include savings achieved since implementation of the technology, efficiency measures identified and implemented through use of the technology, and technology costs.

Twenty-six organizations reported annual energy consumption before and after implementation of the FDD technology. Figure 2 shows the savings results for each participant since the installation of the FDD technology. These savings are based on comparing building energy use in the baseline year prior to FDD implementation, to that in the most recent year for which data were available. These data represent 26 organizations that use FDD across a total 550 buildings comprising 97 million sf of FDD install base. Figure 4 also shows the utility cost savings associated with these energy savings, across the same set of study participants.

The results show that energy savings ranged from -1%-31% percent, with a median of 8%. The median utility cost savings was \$0.27/sf with a range of -0.06-1.3 \$/sf. It is important to note that these savings are not solely attributable to use of the FDD technology, as the FDD was often one component of a multifaceted energy management program, and efficiency measures (e.g. retrofits) not related to use of the FDD were likely implemented during the analysis period. The FDD technology is however, a critical component of respondents' energy management process, and a means of achieving persistence in savings. Among the 26 organizations, nine organizations are in the higher education market sector, eight organizations are in the office market sector, and the remain organizations are in the healthcare, retail, hospitality, and laboratory market sectors. The median energy savings of the higher education organizations and the office organizations are 12% and 8% respectively. The savings results also show that a larger portfolio size is not associated with a greater energy savings percentage.

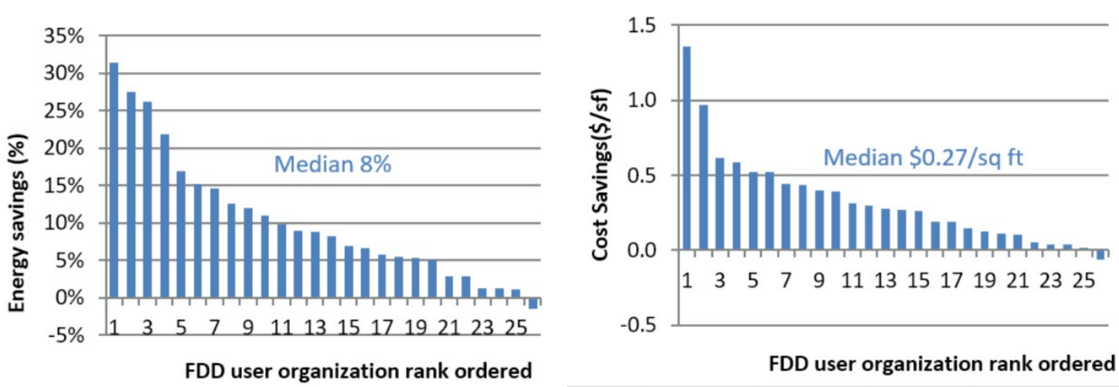

Figure 2: Participant energy savings (left) and cost savings (right) since installation of the FDD technology (n=26)

Study participants were asked to indicate up to 10 of the most frequently implemented measures that were identified through the use of FDD technology, choosing from a list of 26 common operational improvement opportunities. The results are shown in Figure 3, and are consistent with measures commonly implemented in the commissioning process (Mills 2011).

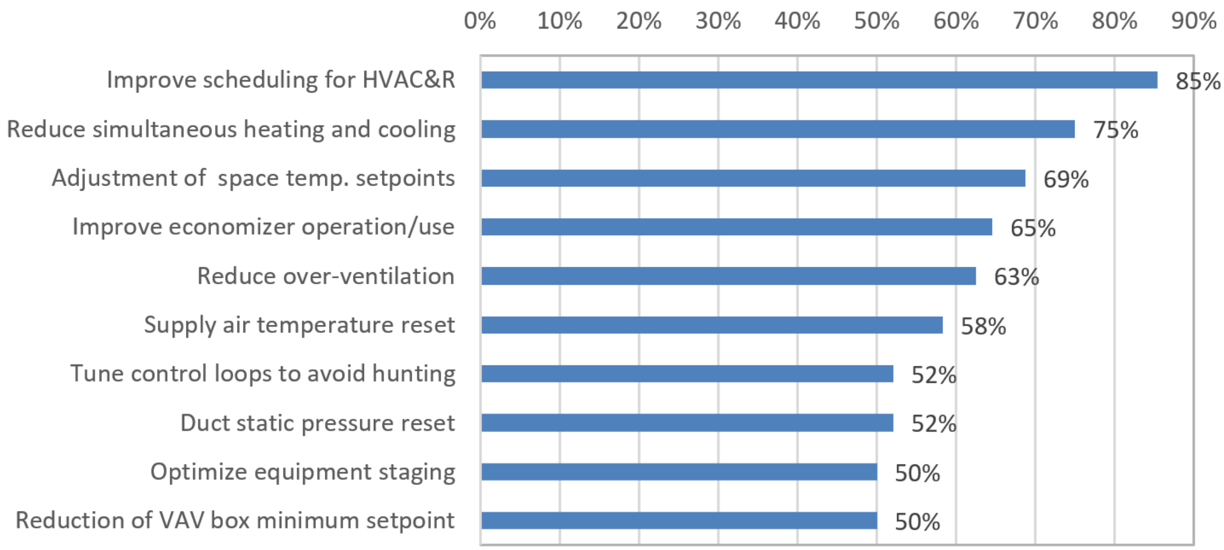

Figure 3: Measures identified and implemented through use of FDD technology (n=26)

Table 3. Ranges and median values of FDD base costs, annual recurring software costs, and annual labor costs

Type of Costs		Costs (N=27)			
		[$]	[$/pt]	[$/building]	[$/sf]
Base Cost	Range	8,000 to 5,000,000	1.1 to 263	1,300 - 83,000	0.004-0.48
	Median	110,000	8	12,500	0.05
Annual Recurring Software Cost	Range	4,000 to 1,600,000	0.3-72	80 – 65,000	0.001-0.16
	Median	33,000	2.7	4,000	0.02
Annual Labor Cost (internal staff or contracted)	Range	9,000 to 5,100,000	0.3 -255	270 – 850,000	0.01-0.85
	Median	60,000-	8	15,000	0.05

Table 3 summarizes the ranges and median of base cost, annual recurring software cost and annual labor cost (internal staff or contracted) across 27 organizations using FDD. Four cost metrics are provided, including total dollars, dollars per data point monitored, dollars per building, and dollars per square feet. Across all cases, the number of points hosted within the FDD ranged from 300 to 200,000, and the median was 1,300 points; the number of buildings in FDD install base ranged from 1 to 1400, and the median was 6 buildings; the size of FDD install base ranged from 0.2 to 52 million square feet, and the median was 2 million square feet.

The median base cost for FDD software installation and configuration was $0.05/sq ft $110,000 total, $8/pt, $12,500/building), and the median annual recurring software cost was $0.02/sq ft ($33,000 total, $2.7/pt, $4,000/building). The median annual labor cost (internal staff or contracted) was $0.05/sq ft($60,000 total, $8/pt, $15,000/building).

3.2 Trial FDD algorithm evaluation methodology and test dataset

3.2.1 Process of implementing evaluation methodology on three algorithms

To assess the ability to execute the FDD algorithm performance testing methodology, input scenarios and daily input samples (steps 1 and 2 in Figure 1) were created from the data summarized in Table 2. To complete step 3 of the process, for each input sample, a *condition-based* convention was used to define the ground truth (faulted or fault free operational state). Detailed in Frank et al. 2018, a condition-based convention defines a fault as the presence of an improper or undesired physical condition in a system or piece of equipment, for example, a stuck damper, or a leaking valve. This is in contrast to behavior-based (e.g. simultaneous heating and cooling) or outcome-based (excessive cooling energy use) fault definitions.

Step 4, running the FDD tools to generate outputs for each input sample, was conducted in two ways, given the two algorithm types that were used in the trial. For the commercial FDD offerings, vendors were provided the input sample data (not ground truth), and information on the system configurations and control sequences. They ran their algorithms against the data, using default thresholds for the fault detection logic, and provided the research team login access to review the FDD results. For the APAR instantiation, the process was simplified since the FDD rules were codified by the authors.

Step 5 in the process entails retrieving the FDD tool outputs for comparison with ground truth. This required manually navigating different elements in the FDD software interface to merge the fault detection and diagnosis outputs for each day, and to identify the tool-generated diagnoses. An example of the outputs from the APAR instantiation and the two commercial FDD offerings, for the same ground truth {*faulted, leaking heating valve*} is shown in Tables 4-6.

The final step of the process entails evaluation of performance metrics by comparing the FDD tool outputs to ground truth, and aggregating across all input samples in the dataset. For the examples shown in Tables 4-6, the ground truth was {*faulted, leaking heating valve*}. Since the APAR instantiation algorithm and both software offerings identified the presence of a fault, the outputs for both were deemed *true positive*. For each FDD algorithm/tool tested, multiple diagnoses were returned. The diagnosis was deemed *correct diagnosis* if one of the listed diagnoses mapped to the ground truth. In the examples in Tables 4-6, there is no diagnosis information for APAR instantiation algorithm, while the diagnosis for both offerings was deemed correct since *leaking heating valve* was named among the potential diagnoses.

Table 4. Fault detection and diagnosis outputs from APAR instantiation

Dataset	Date	Detection Output	Diagnosis Output
MZVAV AHU-2 (Sim.)	8/28/07	Persistent supply air temp error exists (Rule 25)	None

Table 5. Fault detection and diagnosis outputs from an FDD offering 1

Dataset	Date	Detection Output	Diagnosis Output
MZVAV AHU-2 (Sim.)	8/30/07	Supply air temperature higher than setpoint	Simultaneous heating and cooling Undersized coils Stuck or broken dampers or valves Broken or uncalibrated sensor Error in control sequences
		Possible simultaneous or excess heating and cooling	**<u>Valve is not seating properly and is leaking</u>** Stuck or broken valve Temperature sensor error or sensor installation error is causing improper control of the valves or other coils
		Supply static pressure not tracking setpoint	Fan speed control error Damper malfunction Fan malfunction or failure Uncalibrated or malfunctioning pressure sensor

Table 6. Fault detection and diagnosis outputs from an FDD offering 2

Dataset	Date	Detection Output	Diagnosis Output
MZVAV AHU-2 (Sim.)	8/30/07	Under-economizing and cooling	The AHU is using mechanical cooling and not fully utilizing the economizer while outside air temperature is less than return air temperature. Please review the economizer sequence and that the outside air damper is working properly
		<u>Leaking heating valve</u>	**<u>Leaking heating valve</u>**
		Cooling setpoint not met	NA
		Duct static pressure setpoint not met	Confirm the supply fan is not overridden and the setpoint is reasonable for the facility
		Supply air temperature hunting	NA

3.2.2 FDD Algorithm Evaluation Metrics

Figure 4 summarizes the results for each of the detection and diagnosis accuracy metrics that are computed in the test procedure, when aggregated across each input sample in the trial dataset. The dataset includes 47 faulted input samples and 28 un-faulted input samples. The evaluation metrics were calculated following the equations (1) – (7) in the Methodology section. For the current data set, across the three algorithms tested, over half of the faulted samples were correctly detected, with the true positive rate ranging from 70% to 94%. The false positive rates ranged from 36% to 86%, while 26% or less of the faulted samples were missed (false negatives). The true negative rate ranged from 11% to 57%. Only the APAR instantiation was unable to provide a detection result, with a no detection rate of 4%. The algorithm that has the highest true positive rate is also the one with the highest false positive rate. This is not a surprise since higher sensitivity to detect faults can also result in incorrect results when faults are *not* actually present.

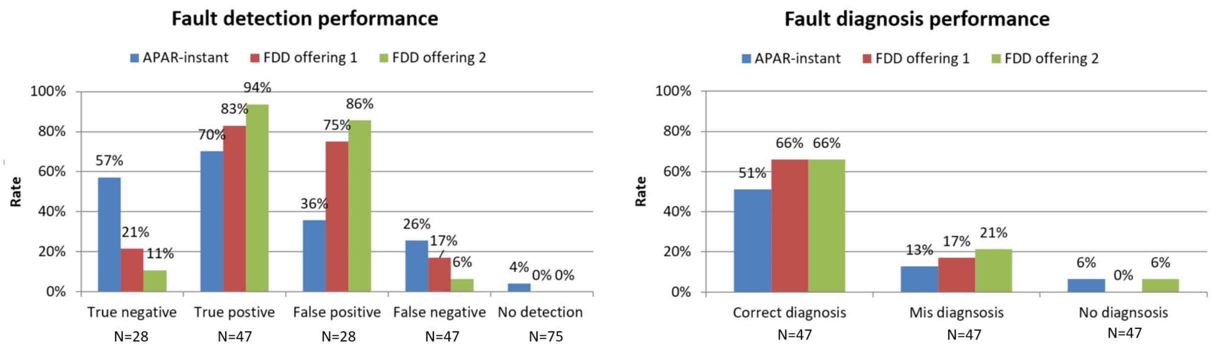

Figure 4. Summary of results from exercising the test procedure and initial dataset against three FDD algorithms (N is the number of observations from which these percentages were calculated).

In addition to detection accuracy, Figure 4 summarizes metrics for fault diagnosis, *following correct detection*. The correct diagnosis rates ranged from 51% to 66%, the misdiagnosis rate ranged from 13% to 21%, and the no diagnosis rate ranged from 0% to 6%.

Across the algorithms tested, there were no significant differences in performance for smaller versus larger fault intensities. With this initial limited dataset, there were also no consistent trends as to which fault types were most likely to be correctly detected and correctly diagnosed. Having confirmed the ability to reliably create valid test data, and execute the performance testing across diverse algorithms, the dataset will be expanded in future work. This expansion will focus on generating a wider range of fault types and intensities under more diverse operational conditions. The expanded data set is expected to facilitate more definitive conclusions to compare algorithms to one another, and derive insights regarding which fault types and intensities are most challenging to detect and diagnose.

In general, the results indicate that the commercial FDD products performed better than the instantiation of the NIST rule set, suggesting that vendors have improved the state of the art since the NIST rules were published in the early 2000s. Nonetheless, it is important to emphasize that this work is meant to illustrate methods that *could in the future* be used to evaluate and benchmark FDD algorithm performance. To draw conclusions about the general performance of FDD technology, or relative performance of one offering versus another, it will be necessary to further expand the current dataset.

4. Discussion

In the first portion of this work, the analysis of a large install base of FDD technology spanning 200 million sf and 36 organizations showed that users are deriving significant benefits, and are doing so cost effectively. Since installation of the FDD technology, users have achieved median whole-building savings of 8%. This is general agreement with prior literature that indicates *potential* savings from FDD of 5-30%. These savings are not *solely* attributable to use of the FDD technology, which may be just one component of a multi-faceted energy management program. However, the FDD technology inarguably plays a key role in identifying operational savings opportunities, and maintaining persistence in those savings over time. The most common measures identified through use of the FDD technology include scheduling, reducing simultaneous heating and cooling, and various controls-related issues related to setpoints, resets, and other problems. These measures represent strong overlap with those commonly found in existing building commissioning, which has been documented to produce median whole building savings of 16% (Mills 2011). In this study, data were collected from the portfolio-level energy managers. In the future, it will be beneficial to survey building-level operators asking about their satisfaction and use of FDD tools as well as the organizational operations and maintenance processes that they implement to respond to the FDD analysis results.

FDD technology pricing models vary, and the costs can be represented in dollars, or normalized by number of buildings or square feet served, or by the number of data points that are continuously accessed, stored, and analyzed in the system. For the purpose of this study, costs were broken into three categories including base cost (median $12.5K per building), annual recurring software cost (median $4K per building), and annual recurring labor cost (median $15K per building). For 21 out of 26 participants, the annual energy cost savings ($/st, Figure 2) exceed median annual recurring costs (software + internal labor) $0.07/sf. The magnitude of these costs can be considered in the context of utility expenditures. The median FDD install base in this study was 2M square feet, and the Energy Information Administration (US EIA 2019) reports that buildings of this size spend an average of nearly $2M annually on energy. When considering the price of FDD software it is important to take into account the full picture of base and recurring costs. For example, with the study cohort, there are instances where the base costs are low but the recurring costs are much higher than average. There are also instances where the base cost is high but there is little to no recurring cost, as the software is hosted and managed in-house.

The second portion of this study demonstrated a trial to implement for AHU systems, an FDD algorithm performance test using two commercial FDD tools and a research-grade FDD algorithm. The results show that the methodology can indeed be executed in a consistent manner across diverse FDD offerings. However, the process is quite manual, and for scalability and repeatability, additional infrastructure to support automation of key steps in the process would be useful. For example, Yuill 2019 offers a platform from which developers can access test data, submit their algorithm's detection and diagnosis outputs, and retrieve a platform-computed set of performance metrics. Much of the manual nature of the process is due to the fact that FDD results are presented differently in different tools, as was illustrated in Section 3.2 Tables 4-6. An important aspect of the testing is therefore to transparently document and define how the FDD tool outputs are matched to the ground truth representation in the test dataset.

The results of the trial evaluation surfaced the fact that FDD routines often return several potential diagnoses. Therefore, it is necessary to define what will qualify as a correct diagnosis. In this trial, a liberal interpretation was used, such that one correct diagnosis amongst a set of possible diagnoses was deemed *correct*. Alternate approaches could apply a weighting factor that accounts for the number of potential diagnoses that are provided. Basic performance metrics (e.g. true positive rate, true negative rate, correct diagnosis rate) were used in this the trial evaluation. There is an opportunity to further consider presentation of evaluation results by fault type when the dataset is expanded to cover more fault intensities and operational conditions.

It is clear that for maximum usefulness the dataset must be expanded. Considering AHU systems, the current dataset is relatively limited in terms of fault intensities and seasonal diversity. Including a wider range of fault types and intensities under a more diverse operational conditions will be needed to more conclusively and thoroughly assess the performance of FDD algorithms. Although low fault intensities can be helpful in teasing out differences in performance between one algorithm and another, it is higher fault intensities that are likely to result in significant fault impacts that would be important to users of FDD technology. Accordingly, the fault impact ratio (Yuill and Braun 2013, 2016) is a complementary metric that could be added to the dataset and presentation of results. Additionally, since today's FDD technologies cover many equipment types and configurations (Granderson et al. 2017b) and often feature libraries of hundreds of diagnostic routines, extending the dataset to cover additional system types would be beneficial.

5. Conclusions and Future Work

Leveraging data from a large study cohort, this paper documented the costs of modern FDD technology, and the technology's role in enabling persistent energy savings over time. It also presented a trial demonstration of how FDD algorithms might be performance tested and benchmarked. Several opportunities for future work are suggested by the findings. First, the study cohort will be expanded, and additional data will be collected to increase the sample size from which as-operated technology benefits can be quantified. This will provide the research and development community as well as industry, the largest available compendium of cost-benefit data for real-world FDD installations, compiled using a consistent and transparently documented experimental design.

The state of today's FDD technology can be advanced through research focused on enhanced diagnostic (as opposed to detection) approaches, and methods for fault prioritization. Complementary work to characterize fault prevalence based on empirical data from the field could also prove valuable in guiding future FDD technology development and implementation efforts.

With respect to FDD algorithm evaluation and performance benchmarking, future work will focus on expansion of the test dataset, and provision of the dataset for public use by FDD research and development community. Although an ambitious undertaking, this would be most useful if informed by findings from a field study on fault prevalence to specify distributions of the data represented in the dataset. Standardization of the fault categories (names and how they are defined), as well as the diagnostic messages would streamline the processes of evaluating different tools or algorithms, however would require buy-in and agreement from a broad ecosystem of developers and product providers. Finally, the standards community and FDD vendor community will be engaged to determine the potential for longer-term formalization of these approaches into standard methods of test, guidelines, standards, or technology certifications.

Acknowledgement

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors wish to acknowledge Marina Sofos and Amy Jiron for their guidance and support of the research, as well as our partners who contributed to the performance testing dataset. Stephen Frank and Xin Jin were instrumental in defining the parameters used in the performance evaluation methodology. Finally, we thank the fault detection and diagnostics technology and service providers who participated in this study.

References

ASHRAE. 2014. ANSI/ASHRAE 140—2014 Standard method of test for the evaluation of building energy analysis computer programs. American Society of Heating Refrigeration and Air Conditioning Engineers,

Bonvini M, Sohn M D, Granderson Wetter M, Piette M A. 2014. Robust on-line fault detection diagnosis for HVAC components based on nonlinear state es- timation techniques, Apply Energy 124:156–166, doi: 10.1016/j.apenergy. 2014.03.009.

Dexter A, Pakanen J (Eds.). 2011. Demonstrating automated fault detection and diagnosis methods in real buildings. 461 Technical Research Centre of Finland, Finland.

Deshmukh S, Glicksman L, Norford L. 2018. Case study results: fault detection in air-handling units in buildings. Advances in Building Energy Research.pp:1-7.

Efficiency Valuation Organization (EVO). 2016. Core concepts: International performance measurement and verification protocol. EVO 10000-1:2016.

Energy Technologies Area (ETA). 2019. Berkeley Lab. FLEXLAB [Internet]. Energy Technologies Area, [cited January 2, 2019]. Available from: https://flexlab.lbl.gov.

ENERGY STAR®. US Environmental Protection Agency and US Department of Energy. [Cited June 14, 2019]. Available from http://www.energystar.gov

Fernandez N, Xie Y, Katipamula S, Zhao M, Wang W, Corbin C. 2017. Impacts on commercial building controls on energy savings and peak load reduction. Pacific Northwest National Laboratory, 2017. PNNL Report #PNNL-25985.

Frank S, Lin G, Jin X, Singla R, Farthing A, Granderson J. 2018. A performance evaluation framework for building fault detection and diagnosis algorithms. Energy and Buildings 192:84-92. DOI: https://doi.org/10.1016/j.enbuild.2019.03.024.

Frank S, Lin G, Jin X, Farthing A, Zhang L, Granderson J. Metrics and methods to assess building fault detection and diagnosis tools. National Renewable Energy Laboratory, February 2019, NREL/TP-5500-72801.

Granderson J, Fernandes S, Singla R, Touzani S. 2017. Corporate delivery of a global smart buildings program. Energy Engineering 115(1):7-25.

Granderson J, Lin G. 2016. Building energy information systems: Synthesis of costs, savings, and best-practice uses. Energy Efficiency 9(6):1369-1384.

Granderson J, Price PN. 2014. Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models. Energy 66: 981-990. DOI: 10.1016/j.energy.2014.01.074.

Granderson J, Price PN, Jump D, Addy N, Sohn MD. 2015. Automated measurement and verification: Performance of public domain whole-building electric baseline models. Applied Energy 144:106-113.

Granderson J, Singla R, Mayhorn E, Ehrlich P, Vrabie D, Frank S. 2017. Characterization and survey of automated fault detection and diagnostic tools. Lawrence Berkeley National Laboratory, November 2017. LBNL Report #2001075.

House JM, Vaezi-Nejad H, Whitcomb JM. 2001. An expert rule set for fault detection in air-handling units. ASHRAE Transactions 107:858-871.

Katipamula S, Brambley M. 2005. Methods for fault detection, diagnostics, and prognostics for building systems – a review, part 1. HVAC&R Research 11(1):3–25.

Kim W, Katipamula S. 2018. A Review of Fault Detection and Diagnostics Methods for Building Systems, Science and Technology for the Built Environment 24 (1) (2018) 3-21.

Lawrence Berkeley National Laboratory (LBNL). Modelica Buildings Library [Internet]. Regents of the University of California; 2018 [cited January 3, 2019]. Available from: https://simulationresearch.lbl.gov/modelica.

Lee KP, Wu BH, Peng SL. 2019. Deep-learning-based fault detection and diagnosis of air-handling units. Building and Environment. 2019 Jun 15,157: 24-33.

Mills E. 2011. Building commissioning: A golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. Energy Efficiency 4(2):145–173. DOI: https://doi.org/10.1007/s12053-011-9116-8.

Muller T, Rehault N, Rist T. 2013. A qualitative modeling approach for fault detection and diagnosis on HVAC systems, in: Proceedings of the 13th International Conference for Enhanced Building Operations, Montreal, Canada

Ferretti NM, Galler MA, Bushby ST, Choinière D. 2015. Evaluating the performance of Diagnostic Agent for Building Operation (DABO) and HVAC-Cx tools using the Virtual Cybernetic Building Testbed. Science and Technology for the Built Environment, 21(8):154-1164. DOI: 10.1080/23744731.2015.1077670.

Roth KW, Westphalen D, Feng MY, Llana P, Quartararo L. 2005. Energy impact of commercial building controls and performance diagnostics: Market characterization, energy impact of building raults and energy savings potential. Report prepared by TIAC LLC for the U.S. Department of Energy.

Ross TM, Braun JE. 1997. A statistical, rule-based fault detection and diagnostic method for vapor compression air conditioners. HVAC&R Research. 3(1):19–37. DOI: 10.1080/10789669.1997.10391359.

Smart Energy Analytics Campaign. Find a Product or Service [Internet]. 2019 [accessed on June 6, 2019]. Available from: https://smart-energy-analytics.org/product-service.

Smart Energy Analytics Campaign. Smart Energy Analytics Campaign [Internet]. 2019 [cited June 14, 2019]. Available from: https://smart-energy-analytics.org.

Summer H, Hilger C. 2012. Fault detection and Diagnostic Software. Pacific Gas and Electric's emerging technology program report ET11PGE3131.

Sun B., Luh P B, Jia Q S, O'Neill Z, Song F. 2014. Building energy doctors: an SPC and kalman filter-based method for system-level fault detection in HVAC sys- tems, IEEE Transaction Automation Science and Engineering. 11 (1): 215–229. DOI: 10.1109/TASE. 2012.2226155 .

Katipamula S, Pratt RG, Chassin DP, Taylor ZT, Gowri K, Brambley MR. 1999. Automated fault detection and diagnostics for outdoor-air ventilation systems and economizers: methodology and results from field testing. ASHRAE Transactions 105(1).

US Department of Energy (DOE) OpenEI. Data Sets for Evaluation of Building Fault Detection and Diagnostics Algorithms [Internet], [cited June 14, 2019]. Available from: https://openei.org/doe-opendata/dataset/data-sets-for-evaluation-of-building-fault-detection-and-diagnostics-algorithms.

US Energy Information Administration (EIA). Commercial Buildings Energy Use Consumption Survey [Internet], [cited June 14, 2019]. Available from: https://www.eia.gov/consumption/commercial/.

Wall J, Guo Y. 2018. Evaluation of Next-Generation Automated Fault Detection & Diagnostics (FDD) Tools for Commercial Building Energy Efficiency – Final Report Part I: FDD Case Studies in Australia, RP1026, Low Carbon Living CRC, pp 68.

Wen J, Li S. 2011. Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Units. Technical Report 1312-RP, ASHRAE.

Yuill D. FDD Evaluator [Internet], [cited January 15, 2019], Purdue University. Available from: https://www.purdue.edu/fddevaluator.

Yuill DP, Braun JE. 2013. Evaluating the performance of fault detection and diagnostics protocols applied to air-cooled unitary air-conditioning equipment. HVAC&R Research 19(7):882–891. DOI: https://doi.org/10.1080/10789669.2013.808135.

Yuill DP, Braun JE. 2016. Effect of the distribution of faults and operating conditions on AFDD performance evaluations. Applied Thermal Engineering 106:1329–1336. DOI: http://dx.doi.org/10.1016/j.applthermaleng.2016.06.149.

Yuill DP, Braun JE. 2017. A figure of merit for overall performance and value of AFDD tools. International Journal of Refrigeration 74: 651–661. DOI: 10.1016/j.ijrefrig.2016.11.015.

Zhao Y, Wen J, Wang S. 2015. Diagnostic Bayesian networks for diagnosing air handling units faults–Part II: Faults in coils and sensors. Applied Thermal Engineering 90: 145-157

Zhao Y, Wen J, Xiao F, Yang X, Wang S. 2017. Diagnostic Bayesian networks for diagnosing air handling units faults–part I: Faults in dampers, fans, filters and sensors. Applied Thermal Engineering 111: 1272-1286.

Zogg D, Shafai E, Geering H P. 2006. Fault diagnosis for heat pumps with parameter identification and clustering, Control Engineering Practice. 14 (12): 1435–14 4 4, DOI: 10.1016/j.conengprac.20 05.11.0 02.